COMPARISON OF TWO DIFFERENT TEXT-TO-SPEECH ALIGNMENT SYSTEMS: SPEECH SYNTHESIS BASED VS. HYBRID HMM/ANN

O. Deroo, F. Malfrere and T. Dutoit
Dept. of Circuits Theory and Signal Processing,
Facult Polytechnique de Mons, boulevard Dolez 31,
Mons 7000, Belgium
Tel: +32 65 374133; fax: +32 65 374129
e-mail: {malfrere,deroo,dutoit}@tcts.fpms.ac.be

ABSTRACT

In this paper we compare two different methods for phonetically labeling a speech database. The first approach is based on the alignment of the speech signal on a high quality synthetic speech pattern, and the second one uses a hybrid HMM/ANN system. Both systems have been evaluated on French read utterances from a speaker never seen in the training stage of the HMM/ANN system and manually segmented. This study outlines the advantages and drawbacks of both methods. The high quality speech synthetic system has the great advantage that no training stage is needed, while the classical HMM/ANN system easily allows multiple phonetic transcriptions. We deduce a method for the automatic constitution of phonetically labeled speech databases based on using the synthetic speech segmentation tool to bootstrap the training process of our hybrid HMM/ANN system. The importance of such segmentation tools will be a key point for the development of improved speech synthesis and recognition systems.

1 INTRODUCTION

The use of corpus-based methods, where knowledge is automatically derived from large speech corpora, has become the primary methodology in the areas of text-to-speech synthesis and speech recognition. These machine learning approaches have emerged essentially thanks to the development of more and more powerful computers and computational models of speech, like artificial neural networks (ANNs) or hidden Markov models (HMMs). To use such corpus-based approaches, large speech databases are needed. Most of the time a phonetic transcription aligned with the speech corpora is also required. The labeling of such corpora is very tedious and time consuming and so involves a non negligible cost in the development of speech systems. To reduce the cost and the time needed to label a speech corpus, several authors have proposed various systems based on HMMs [1],[2]. The major drawbacks of such techniques is the need of a training stage, that is, of a training database. Another way of performing such an alignment is to use a speech synthesizer as described in [3], which offers the great advantage that no training stage is needed to perform the alignment. This paper compares the results, the advantages and the drawbacks of these two radically different approaches of the phonetic speech segmentation problem. The paper is divided as follows. In section 2, the first system based on the use of a speech synthesizer is described. Section 3 details the hybrids HMM/ANN approach used for the phonetic alignment. Section 4 give a comparison of the results obtained with both methods on the same speech corpus. The article ends with some conclusions and comments about the two approaches in section 5.

2 SPEECH SYNTHESIS BASED PHONETIC ALIGNMENT

The main idea of the speech synthesis based phonetic alignment is to use a digital speech synthesizer to create a reference speech pattern with a predetermined phonetic segmentation and then align the natural speech on this pattern. Figure 1 shows the steps needed to implement a text-to-speech alignment system based on speech synthesis techniques. In a first stage the phonetic transcription is automatically derived from the text with an accurate automatic phonetization system like those used in text-to-speech synthesis systems. The publicly available speech synthesizer MBROLA [4], which is based on diphone concatenation techniques, is then used to generate a reference speech signal from the phonetic transcription. Although natural prosodic information is needed to deliver natural sounding synthetic speech, a very rough prosody suffices to obtain the reference signal since only its segmental features will be used during the temporal alignment process. Phoneme duration and intonation contours are chosen so as to facilitate the alignment process. Phoneme duration are correlated with the local continuity constraint of the alignment algorithm. A constant duration of a hundred milliseconds has been chosen [5]. Since no assumption can be made

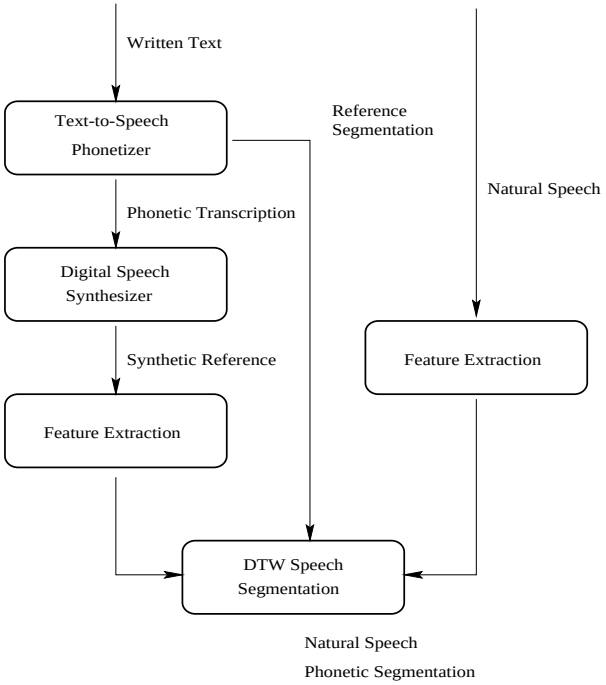

Figure 1: Text-to-speech alignment system.

on the contour actually produced by the speaker, the synthetic F0 curve is chosen as simple as possible (constant F0 value). Assuming the features used to compare the reference and the test signals are not correlated with the F0 curve, this choice has no important effect on the accuracy of the segmentation. To compare the synthetic reference speech and the original speech, some relevant features must be extracted from both signals. Four set of parameters have been used to characterize speech frames:

- the 18 first cepstral coefficients (c_i) derived from a linear prediction analysis (12th order). These coefficient are normalized (CMS) and weighted with a sinusoidal function;
- the temporal derivative cepstral coefficients (Δc_i) are computed in order to account for their time variation;
- the normalized energy (E) of each frame;
- the delta energy (ΔE);

The resulting 38 coefficients are known to result in a good representation of the local spectral envelope. Finally, the segmentation process takes place. It is based on a classical dynamic time warping (DTW) algorithm based on the minimization of the accumulated distance between the two speech signals. The distance used to compare a frame of the synthetic reference and a frame of the input speech is a weighted combination of several euclidian distances: the cepstral distance and an energy distance:

$$\begin{aligned} d(a,b) &= \alpha \sum_{i=0}^{18} (c_i(a) - c_i(b))^2 \\ &\quad + \beta \sum_{i=0}^{18} (\Delta c_i(a) - \Delta c_i(b))^2 \\ &\quad + \gamma (E(a) - E(b))^2 \\ &\quad + \varphi (\Delta E(a) - \Delta E(b))^2 \end{aligned} \quad (1)$$

The great advantage of this approach is that there is no training stage, that is no training database is needed. As a result, the system can be easily adapted to align different languages. Segmentation results for English, German, Dutch, French, Spanish and Romanian can be found in [3]. In comparison with HMMs, the approach apparently looses speaker independence (only one voice is used as reference). However, we found in [3] that it is not the case in practice. This system has been integrated in a prosody transplantation tool called MBROLIGN. It can be freely downloaded for academic purposes from a our Web site http://tcts.fpms.ac.be/synthesis/mbrolign .

3 HYBRID HMM/ANN PHONETIC ALIGNMENT

The hybrid HMM/ANN system used to align the data has been trained on BREF-80 [6]. BREF is a large speech corpus extracted from the French newspaper Le Monde read by 80 speakers. This database is the training material used for our baseline hybrid HMM/ANN system. In all cases we used embedded Viterbi training, a procedure which requires a phonetic labelling of the database. As no phonetic segmentation is provided with BREF, we generated a first segmentation using the method introduced in Section 2 [3]. This first segmentation was used to bootstrap the training and segmentation procedures. First, a multi-gaussian HMM system was trained and after some iterations a quite accurate segmentation was obtained for the whole training set. The multi-gaussian system used diagonal covariance matrices and the number of gaussians per state was chosen equal to 16. This HMM system has been used to generate the segmentation of the whole training set which was used to train a Neural Network (Multilayer Perceptron). The training set of the BREF corpus consists of 3737 sentences (3363 for training and 374 for cross validation) from 56 speakers. We defined a small test set composed of 144 sentences from 8 speakers (4 females and 4 males) in order to check the accuracy of our hybrid HMM/ANN system. Three sets of acoustic features have been used : the log-RASTA-PLP and PLP cepstral features [7] and the LPC-cepstral features with cepstral mean subtraction (CMS) [8]. These features have been chosen for their robustness against channel and speaker characteristics. These parameters (com-

		log-RASTA	CMS	PLP
Train	2.400.000	79.6 %	76.3 %	82.0 %
Cross	270.000	77.0 %	74 %	80.2 %
Phone	35	69.2 %	63.8 %	72.9 %

Table 1: Recognition rate at the frame and phone levels using a classical hybrid HMM/ANN trained on log-RASTA-PLP and CMS features.

monly used in speech recognition systems) were computed every 10 ms on 30 ms analysis windows. The LPC analysis order was set to 10. Thus the feature set for our hybrid HMM/ANN systems was based on a 26 dimensional vector composed of the cepstral parameters (log RASTA-PLP or LPC-cepstral parameters with cepstral mean subtraction), the Δcepstral parameters, the Δenergy and the $\Delta\Delta$energy. Nine frames of contextual information are used at the input of the ANN, leading to 234 inputs (9 frames of context being known as yielding usually the best recognition performance [9] [10]). The training and cross-validation scores at the frame level achieved with this system are given in Table 1. It shows a phone recognition rate of 73 % using a set of 35 CI (Context Independent) phone models. No optimization (duration modeling or phone language model) was realized on this particular task. This recognition rate is the best reported in the literature on this particular task using such a simple system. All the experiments reported in this article related to the hybrid HMM/ANN system have been realized with the STRUT [12] software.

4 RESULTS AND COMPARISON

In this section, the baseline system of section 3 is used in order to label utterances never seen in the training data from a speaker and is compared with the one reported in section 2. Both systems have been compared on the same speech corpus, composed of twenty six utterances read by one speaker with an average duration of 13,6 seconds. That corpus totals 6829 phoneme transitions. The speaker was never seen in the training stage of the HMM/ANN system and was not the one who recorded the diphone database of the speech synthesizer. The corpus was manually segmented to allow the comparison of all segmentation results to a reference. In Table 2, the segmentation errors of both systems are ranked as a function of their amplitude and according to the transition type: consonant-consonant (C-C), consonant-vowel (C-V), vowel-consonant (V-C) and vowel-vowel (V-V). Table 2 also gives the composition of the corpus. Notice that experiments [11] have shown that two human labelers may also disagree over more than 20 ms in about 10 % of the cases. Taking this into account, the two systems described here lead to equivalent results. Most errors are encountered on vowel

%	C-C	C-V	V-C	V-V
Corpus	10,97	35,06	37,06	1,38
Hybrid HMM/ANN Based %				
$<$ 10 ms	84,73	79,12	81,05	66,67
$<$ 20 ms	87,93	83,36	83,89	70,59
$<$ 30 ms	91,87	89,37	88,85	82,35
$<$ 40 ms	95,07	92,53	92,35	86,27
$<$ 50 ms	97,04	95,30	95,34	92,16
$>$ 50 ms	2,96	4,70	4,66	7,84
Speech Synthesis Based %				
$<$ 10 ms	66,43	69,62	68,95	50,00
$<$ 20 ms	82,78	81,98	82,51	70,00
$<$ 30 ms	89,10	87,79	87,86	78,00
$<$ 40 ms	93,50	93,02	92,41	86,00
$<$ 50 ms	96,31	95,06	95,39	92,00
$>$ 50 ms	3,69	4,94	4,61	8,00

Table 2: Phonetic Segmentation Results.

to vowel, vowel to nasal consonant, vowel to liquid and silence to plosive transitions in both approaches. The main advantage of the speech synthesis-based approach is that no training stage is needed. On the other hand, HMM based systems can easily take multiple phonetic transcriptions (phonetic lattice) into account, a feature which is needed when the exact phonetic transcription is not known.

5 CONCLUSIONS

The hybrid HMM/ANN system used here produced an accurate segmentation of a database. The alignment is achieved in real-time and one can use a phonetic lattice in order to deal with homophone pronunciations and liaisons (which are often encountered in French). On the other hand, HMMs are quite difficult to train. They need a first segmentation in order to bootstrap the embedded Viterbi training. Although it is possible to use a linear segmentation, the convergence of the training algorithm is greater with a more accurate initial segmentation. The training of our hybrid HMM/ANN system has been bootstrapped with the segmentation obtained with the alignment system described in section 2. This lead to a very good phone recogntion rate for our hybrid system showing that it is possible to have quickly quite good system. Thus the combination of the two methods results in an efficient way of generating accurate segmentation of speech data. It therefore opens new perspectives for the training of speech recognition models in many languages.

6 ACKNOWLEDGMENTS

This work is supported by a F.R.I.A. grant (Fonds pour la Formation à la Recherche dans l'Industrie et dans l'Agriculture).

References

[1] Talkin D. and Wightman C. W., "The Aligner : Text to Speech alignment using Markov models and a pronunciation dictionary", Proceedings of Second ESCA/IEEE Workshop on Speech Synthesis, pp. 89-92, 1994.

[2] Van Coile B., Van Tichelen L., Vostermans A., Wang J. W. and Staessen M., "PROTRAN: A Prosody Transplantation Tool for Text-to-Speech Applications", Proceedings of ICSLP'94,1994.

[3] Malfrere F. and Dutoit T., "High-Quality Speech Synthesis for Phonetic Speech Segmentation", Proceedings of EuroSpeech'97, vol. 5, pp. 2631-2634, 1997.

[4] Dutoit T., Pagel V., Pierret N., Bataille F. and van der Vrecken O., "The MBROLA Project : Towards a Set of High Quality Speech Synthesizers Free of Use for non commercial purposes", Proceedings of ICSLP'96, pp. 1393-1396, 1996. (http://tcts.fpms.ac.be/synthesis)

[5] Malfrere F. and Dutoit T.,"Speech Synthesis for Text-to-Speech Alignment and Prosodic Feature Extraction", Proceedings of ISCAS'97, pp. 2637-2640, 1997.

[6] Lamel L., Gauvain J. L. and Esknazi M., "BREF, a Large Vocabulary Spoken Corpus for French", Proceedings of EuroSpeech'91, 1991.

[7] Hermansky H. and Morgan N., "RASTA processing of speech", IEEE Transactions On Speech and Audio Processing, Vol. 2, n. 4, pp. 578-589, 1994.

[8] Van Hamme H. ,Gallopyn , Van Springel W., D'Hoore B., Butnaru M, and Bourlard H., "Acoustic Features Comparison and Robustness Tests of a Real Time Recognizer on a Hardware Telephone Line Simulator", Proceedings of ICSLP'94, 1994.

[9] Bourlard H. and Morgan N., "Connectionist Speech Recognition A Hybrid Approach.", KLUWER Academic Publisher, 1994.

[10] Deroo O., Ris C., Malfrere F., Leich H., Dupont S., Fontaine V., and Boite J.M., "Hybrid HMM/ANN systems for speaker independent continuous speech recognition in French", Proceedings of PRORISK'97, p. 137-141, 1997.

[11] Cosi P., Falavigna D. and Olmologo M., "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation", Proceedings of EuroSpeech'91, pp. 693-696, 1991.

[12] Boite J.M., Deroo O., Ris C., Fontaine V.,"Step by Step guide to using the Speech Training and Recognition Tool (STRUT)". Available via ftp at http://tcts.fpms.ac.be/speech/strut/users-guide/users-guide.html